Journal of Biomedical Informatics
Manuscript Draft

Manuscript Number: JBI-18-799

Title: Drug-Target Protein Interaction Prediction Method based on DBN

Article Type: Research paper

Keywords: Deep belief network; Drug-target protein; Interaction prediction; Deep learning; Bioinformatics

Abstract: Drugs may have multiple drug targets, and the most of targets are composed of different proteins. Therefore, the study of drug-target interaction (DTI) prediction has important meaning in drug repositioning, drug development time shortening and the cost of drug research and development reducing.
In this paper, we proposed a deep belief network-based DTI prediction algorithm: we extracted extended connected fingerprint of the drug from the molecular structure. And then, we extracted the structure characteristics of the three peptide of the protein from the amino acid sequence of the protein. At last, we train the deep belief network by the characteristic vector extracted from drugs and proteins. In our proposed method, we fully use of the characteristics in the deep learning network and integrate the empirical feature selection into the deep belief network. Compared with the state-of-the-art approaches, the experimental results show that our method outperforms the other algorithms in massive data sets.

Cover Letter

Dear Editors:

We would like to submit the enclosed manuscript entitled “Drug-Target Protein Interaction Prediction Method based on DBN”, which we wish to be considered for publication in “journal of Biomedical Informatics”. No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Copy of the Abstract:

Drugs may have multiple drug targets, and the most of targets are composed of different proteins. Therefore, the study of drug-target interaction (DTI) prediction has important meaning in drug repositioning, drug development time shortening and the cost of drug research and development reducing.

In this paper, we proposed a deep belief network-based DTI prediction algorithm: we extracted extended connected fingerprint of the drug from the molecular structure. And then, we extracted the structure characteristics of the three peptide of the protein from the amino acid sequence of the protein. At last, we train the deep belief network by the characteristic vector extracted from drugs and proteins. In our proposed method, we fully use of the characteristics in the deep learning network and integrate the empirical feature selection into the deep belief network. Compared with the state-of-the-art approaches, the experimental results show that our method outperforms the other algorithms in massive data sets.

The following is a list of possible reviewers for your consideration:

1) Name: Wei Luo E-mail: luowei80@gdut.edu.cn

2) Name: Qing-Feng Zhang E-mail: tqfz@jnu.edu.cn

3) Name: Shou-Bin Dong E-mail: sbdong@scut.edu.cn

We deeply appreciate your consideration of our manuscript, and we look forward to receiving comments from the reviewers. If you have any queries, please don't hesitate to contact me at the address below.

Thank you and best regards.

Yours sincerely,

Yi-Chen He

Corresponding author:

Name: Yi-Chen He

E-mail: heyichen666@hotmail.com

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Drug-Target Protein Interaction Prediction Method based on DBN

GU Wan-Rong [1], He Yi-Chen(Corresponding author) [1], Xie Xian-Fen [2], Zhang Zi-Ye [1], and Li Jia-Lang [1],

[1]School of mathematics and informatices, South China Agricultural University, Guangzhou, Guangdong 510642, China
[2]School of economics, Jinan University, Guangzhou, Guangdong 510632, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Drugs may have multiple drug targets, and the most of targets are composed of different proteins. Therefore, the study of drug-target interaction (DTI) prediction has important meaning in drug repositioning, drug development time shortening and the cost of drug research and development reducing. **Results:** In this paper, we proposed a deep belief network-based DTI prediction algorithm: we extracted extended connected fingerprint of the drug from the molecular structure. And then, we extracted the structure characteristics of the three peptide of the protein from the amino acid sequence of the protein. At last, we train the deep belief network by the characteristic vector extracted from drugs and proteins. In our proposed method, we fully use of the characteristics in the deep learning network and integrate the empirical feature selection into the deep belief network. Compared with the state-of-the-art approaches, the experimental results show that our method outperforms the other algorithms in massive data sets. **Availability:**https://pan.baidu.com/s/1GFRSUWGEb58vRiSJW-PUNg **Contact:** heyichen666@hotmail.com **Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

Prediction of drug-protein interactions is an important research in recent drug effects discovery. Drug discovery research is an inefficient and costly research and development process, with very expensive costs about $1.8 billion for each new molecular entity discovery. Besides, scientist would take about 10 years to make a new drug to reach medical market; for example, only about 20 new drugs have been approved by FDA as new molecular entity every year [CYZ+15]. For the past decade, the drug development has decreased year by year. For the above reasons, the abandoned or existing drugs are very important and urgent for their new use. Such a new use among drug and protein research is called drug repurposing or repositioning.

In the past few decades, many scholars have studied drug-protein interaction systematically. Newly-published studies on drug discovery follow the model of “one molecule, one goal, one disease”. This model identifies that some molecules interact with special proteins. However, the main limitation is that the drug is map to a target protein based on prior experience in mathematics model. In fact, many diseases are very complex with multi-factors, for instances, they would contact with different genes or different pathways. The accuracy of drug discovery would be very easily affected in this multifactorial environment. Traditional studies ignore the complex diseases associated with the complex factors. For the above issues, this research pattern would not good for drug discovery as expected. Most drug targets are cell proteins, which are designed to treat or diagnose disease through selective interactions with the compound. Current papers show that classical drug targets has about 130 protein families. It is estimated that there are about 6,000 ~8,000 pharmacological targets in the human genome. So far, however, only a small number of these targets have validated with approved drugs, and a lot of presumed drug targets still to be verified in future work.

With the deep study of the drug-protein interactions, the number of related public data set has also increased considerably. For example, DrugBank dataset is an annotated cheminformatics and bioinformatics data set that combines detailed drug, protein and their interaction data. It includes drug types, drug profiles, chemical structures, drug composition, drug targets, drug interaction, etc. This database is supported by the Canadian Institute of Health. STITCH is a public well known data set, including drugs, protein, chemical structures and chemical-protein

interactions, which combined the results from drug experiments, other public data sets and published literatures. Newly studies show that, about 96% drugs approved by FDA has resistant protein target. Due to the high dimensionality and complexity of the characteristics of drugs and proteins, recent studies have proposed to use the machine learning methods to mine the interactions between drugs and proteins automatically. By means of numerical analysis and excavation, the problem of drug-protein interaction prediction is transformed into collaborative filtering, matrix decomposition or score prediction, which improves the efficiency of the prediction significantly compares to the manual experiments. However, some proteins remain have off-target effects and even let the drugs show the toxic side effects. Although the use of machine learning methods has been shown to have a clear advantage in the research of drug-protein interaction prediction, there still has at least two problems in this research:

(1) In a single machine learning algorithm, it is difficult to mine the complex parameters and the implicit association. Besides, If the model is too simple, its predictive accuracy will be affected. Conversely, if the model is too complex, it is easy to cause the over-fitting problem.

(2) The cold drug association problem can not be solved very well in the matrix-based predictive algorithm. Cold association problem refers to the phenomenon that some objects have little relation with other objects in the research of recommended algorithm or score prediction. If there is too much cold association, it has a great influence on the prediction algorithm process.

In order to solve the above problems, we propose a more efficient method with improved AdaBoost algorithm, which combines several weak classifiers to form a powerful classifier. Besides, the problem of scoring prediction can be transformed into classification problem in our new model. This algorithm frame is applied to the matrix filling research of drug-protein interaction prediction. The method proposed in this paper makes use of machine learning algorithm to predict the interaction of drug-protein, which has the advantages of high efficiency and low cost. At the same time, the limitations of the single machine learning algorithm have been overcame significantly. It can better explore the hidden factors and improve the accuracy of prediction effectively. In the case of cold drug associations, our model is also better than other approaches. The remaining of this paper is organized as follows. Section 2 covers related work relevant to our study, including network-base model, machine learning-based method and other methods in prediction of drug-protein interaction. In section 3, we show the DTI prediction method based on Deep Confidence Network. Besides, we show the detail of our proposed algorithm framework. Section 4 shows the experiments and results analysis compared to classical models and the state-of-the-art approaches. Finally, we conclude this manuscript and discuss the future work in section 5.

2 Related Works

Traditional DTI prediction methods can be roughly divided into two categories: docking simulation method and ligand-based method [DTMZ13]. Docking simulation method deduces the relationship between drug and protein from the structural information of the drug, the amino acid sequence and structural information of the protein. Ligand-based method mine the potential relationship through the interactions between drug ligands and of target protein ligands. Usually, the accuracy of the former is higher than that of the latter, but the material structure needs to be calculated in docking simulation method. Therefore, it cost a lot of time. In addition, some protein structure information is not suitable for the use of docking simulation, such as G protein coupling receptor. The ligand-based method does not require too much computation on the material structure and has well scalable. However, when the number of drugs and protein ligands is insufficient, the effectiveness of this method is not satisfactory. Similarity-based DTI prediction approach is also a common method. Yamanishi et.al used statistics and digraphs to make DTI predictions, which are classified according to the target protein type of the Kegg DRUG data. Four standard data sets, Enzyme, GPCRs, Icon channel and Nuclear receptors were constructed by Yamanishi et [YAG+08, YKKG10].al to record the drug-protein relationship, drug similarity and protein similarity. These standard data sets are widely used by subsequent DTI prediction research studies. DTI Predictions are similar to recommended technology, allowing for the use of interconnectedness of objects and information carried by objects to predict potential associations or user ratings. Therefore, there are a lot of recent researches using recommendation technology theory to study DTI prediction. DTI prediction studies usually used the well-known databases such as Drugbank, Chembl and Kegg DRUG [GBB+11, Kan02, LKD+13], which record the characteristics of drugs, target protein characteristics and reaction conditions.

2.1 The Research of Recommendation Technology DTI

2.1.1 Collaborative filtering

The collaborative filtering algorithm is one of the common recommendation algorithms. According to the recommended objects, the algorithms can be divided into user-based and item-based collaborative filtering algorithms. In the case of DTI prediction, users and items can be replaced with drugs and proteins.

	t_1	t_2	t_3	t_4	t_5	t_6
d_1	1	0	0	1	0	0
d_2	0	0	0	1	1	1
d_3	0	1	1	0	0	0
d_4	0	1	0	1	1	0
d_5	1	1	0	1	0	1

Fig. 1. Matrix representation of collaborative filtering method.

Set a drug collection as $D = \{d_1, d_2, ..., d_M\}$, a protein collection as $T = \{t_1, t_2, ..., t_N\}$, here, $|D| = M$, $|T| = N$. Then there is a correlation Matrix R, as shown in Fig.1. The size of R is $M \times N$, and $r_{ij} = 1$ if the known drug d_i interacts with the protein t_j, and $r_{ij} = 0$ otherwise. Based on the Association Matrix R, the associated vectors of the drug and protein can be obtained. And then, the similarity of Pierre Carson Coefficient, cosine similarity, Jaccard similarity can be used to calculate the similarity between the drugs and protein. Finally, the potential relationship between d_i and t_j can be predicted due to Eq.(1).

$$r_{ij} = \frac{\sum_{d_t \in top(d_i,k)} similar(d_i, d_t) \times r_{it}}{\sum_{d_t \in top(d_i,k)} similar(d_i, d_t)} \quad (1)$$

2.1.2 Graph Model

The graph model can represent the different entities and their relationships. The relationship of drugs and proteins can also be predicted in a graph model, Their similarity and known relationships can be expressed as side-to-side representations, and then the information of the graph model can be used for DTI prediction, as shown in Fig.2. The data representation based on graph model can express the related information of drug and protein conveniently and intuitively, so there are many DTI prediction research based on graph [APGF13, EIP+13].

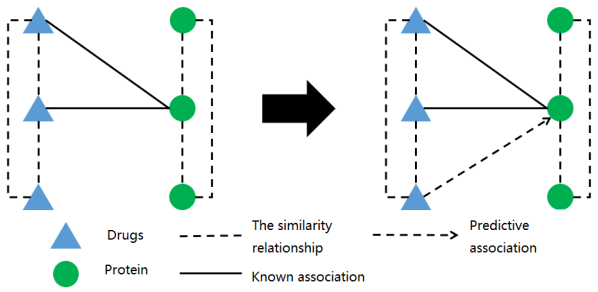


Fig. 2. Graph model based drug-target interaction prediction.

The DTI prediction based on the graph model is usually based on two hypotheses below: (1) the more similar the compound structure between the two drugs, the more likely it is to work on the same target protein; (2) the more similar amino acid structure the two target proteins have, the more likely they are to be affected by the same drug. So in addition to figure 2, the use of side connections indicates the relationship between the drug and the drug, protein, and protein similarity, a more careful alternative to the similarity side is the introduction of two nodes, the molecular structure of the compound and the structure of the protein, which are directly linked to a molecule if the drug and protein have a substructure, this allows for the use of graph theory to calculate their similarity based on the common substructure of the object, thus predicting the potential link between the drug and the protein.

2.1.3 Implicit Semantic Model

The implicit semantic model (LFM) is an improvement on the decomposition of singular matrices. In DTI prediction, the relationship between the drug and the implicit classification, protein and the implicit classification is established by Matrix decomposition. In the Association Matrix R, a lot of relationships are implicit. The aim of LFM is to excavate the hidden factors in the process of Matrix decomposition and to predict the relationship of unknown action.

The loss function of LFM can be expressed as:

$$min||R - \tilde{R}||^2 = min||R - AB^T||^2 \quad (2)$$

$\tilde{R}$ denotes the score prediction matrix. A denotes the relationship matrix between drugs and implied taxonomy. B denotes the association matrix of proteins and implicit taxonomy. $A \in R^{M\times K}$, $B \in R^{N\times K}$. K denotes the implicit taxonomy. Once the loss function is given, you can use the mathematical optimization to minimize the loss function. Ezzat et al. [EZW+17] in GRMF, it uses the KNN Algorithm to calculate the most approximate k-neighborhood elements of each drug and protein, in this paper, the constraint of minimizing the element distance in k-neighborhood is added to the loss function of the follow-up training, which ensures that the characteristic vectors of drugs and proteins obtained by matrix decomposition are similar to those in k-neighborhood.

2.2 The Related Works of Deep Learning in DTI

The commonly used drug targeting protein databases (DrugBank and Chembl) are very sparse in association number, the traditional recommendation algorithm is easy to reach the bottleneck, and the effect of DTI prediction is limited. By using the deep learning method, the characteristics can be obtained automatically according to the network structure and the relation of Sparse Matrix can be explored.

Wan Et Al. [WZ16] use the Morgan's fingerprints and tf-idf Algorithms to obtain a word frequency matrix of the drug and drug substructure, and then use the singular value decomposition to derive the drug signature vector, the amino acid sequence of protein was considered as a word, and the character of protein was calculated by word2vec method. Finally, it formed the input vector of 300. Using A six-layer neural network (600-256-128-64-32-16), the middle transmission function uses ReLU function and uses the dropout module. The experiment results show that this method is superior to the Random Forest Algorithm.

Peng et Al. [CY+16] used only Morgan's fingerprint to represent the drug's characteristics and to use a multiscale protein sequence representation to represent the protein signature, which forms the input vector of length 1448 as input for SAE, and finally, sae encodes it as a feature vector of length 200 and gets the feature vector as input to the SVM classifier. The method uses SAE to compress the original data by coding and extract its depth feature and reduce the dimension, instead of using SVM classification directly. The experimental results show that this method is superior to the network-based, similarity-based method and PSSM Algorithm. The deep neural network has many training parameters which can easily produce over-fitting problems [LHZ17. To solve this problem, Hinton proposed the deep confidence network (DBN), which allows network parameters to be initialized in better and more stable positions through an initialization parameter method based on the energy function, and then fine-tune the network. Hinton showed that the performance of DBN was better than that of the deep neural network connected by hand-written digital recognition. And in subsequent studies [SMH07] the user-rating-data matrix was used as input, a single limited Boltzmann machine was used to estimate the score of other goods and achieved good results.

3 DTI Prediction Method Based on Deep Confidence Network

In this paper, we propose to use deep confidence network to make DTI prediction. The training process of DTI is as follows:

(1) to calculate the extended connection of drug by fingerprint signature and the three-peptide structure characteristics of the amino acid sequence of protein.

(2) pre-training for parameters of each layer of deep confidence network based on restricted Boltzmann machine (see section 2.1)

(3) fine-tuning the network using back propagation (see section 2.2 for details).

3.1 Parametric Training for Deep Trust Networks

The restricted Boltzmann machine (RBM) is based on the theory of the Boltzmann machine (BM). BM is a binary Markov random field, which can be represented by a weighted, directionless graph, as shown in figure 3(left), each node can be connected at will. In order to reduce the complexity of the model and reduce the training difficulty and time complexity of the model, RBM limits the connection of nodes. As shown in figure 3(right), nodes at the same layer in RBM are not interconnected, RBM can be divided into visible layer and hidden layer according to the hierarchy.

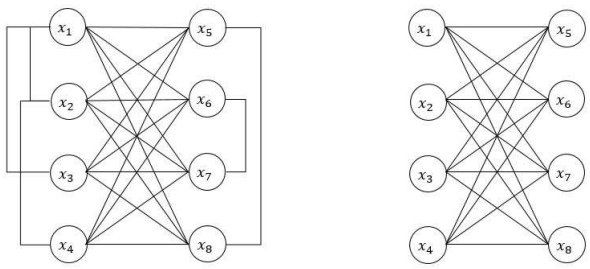


Fig. 3. Boltzmann aircraft (left) and restricted Boltzmann machine (right).

Each node in the RBM is binary, that is, it has 2_n states. When $x = 1$ indicates that the current node is open, and the node is close otherwise. The current RBM energy values can be calculated using the following energy functions:

$$E(X) = \sum_{i=1}^{n} b_i x_i - \sum_{i,j} xiw_{i,j}x_j \tag{3}$$

where b is the offset value of node i, $w_{i,j}$ is the weight of the connection edges of node i and j. Energy functions are a measure of the state of the system as a whole, and the more orderly or distributed the system is, the less energy the system has. On the contrary, the more disordered the system, the greater the energy. RBM is a kind of unsupervised learning model. The purpose of RBM is to make the RBM fit the distribution of input data to the maximum extent possible. According to the energy function, the probability distribution of x in different states can be calculated under given parameters:

$$P(X) = \frac{1}{Z}e^{(-E(X))} \tag{4}$$

$$Z = \sum_{X} e^{(-E(X))} \tag{5}$$

In both formulas above, Z is the normalization constant. The RBM model is divided into the visible layer and the hidden layer. The RBM model used in this paper is shown by the dotted line box in figure 4. RBM has data encoding ability, V denotes visible layer, H denotes hidden layer, V' denotes encoded state of H after decoding, and the number of nodes is n_V, n_H, $n_{V'}$, and $n_V = n_{V'}$.

RBM uses input data to guide its training, first encoding the V as H, and then decoding the H to V'. The goal is to fit V and V' as much as possible. The difference between the two is measured by loss function $L_{W,a,b}(V, V')$. The noise of the data processed by RBM will be reduced, making it easier to get key feature information.

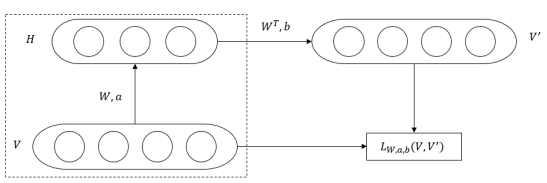


Fig. 4. RBM training model.

In the state of a given visible layer node, the probability distribution of the state of the hidden layer node can be obtained. Similarly, the probability distribution of the visible layer can be obtained by giving the node state of the hidden layer:

$$p(h_j = 1|V) = \sigma(a_j + \sum_{i=1}^{n_V} v_i w_{ij}) \tag{6}$$

$$p(v_i = 1|H) = \sigma(b_i + \sum_{j=1}^{n_H} h_j w_{ij}) \tag{7}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. The above two formulas get the probability that each node state in H or V is an open state (when $h_j = 1$ or $v_i = 1$). In order to train the RBM model, the node state needs to be sampled according to the probability distribution of the node state.

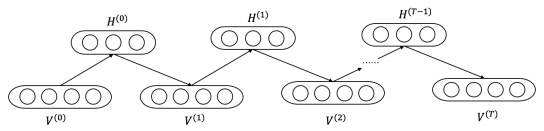


Fig. 5. T-step Gibbs Sampling IN RBM.

In general, the RBM model sampling method uses Gibbs sampling, as shown in Figure 5. RBM's Gibbs sampling process is as follows:

Algorithm 1 Sampling algorithm flow

Require: The state value of the visible Layer V; Number of iterations T;
Set $V = (v_1, v_2, v_3, ..., v_{n_V})$, $H = (h_1, h_2, h_3, ..., h_{n_H})$.
Ensure: After t-iteration, the state value of V
(1) Initialize the iterations $t = 0$ and make $V^{(0)} = V$;
(2) For each node of the hidden layer H: $h_j^{(t)}$, $j = 1, 2, ..., n_H$
$h_j^{(t)} \sim p(h_j^{(t)} = 1|V^{(t)})$;
(3) Set $t = t + 1$;
(4) For each node of the visible Layer $V^{(T)}$: $v_i^{(t)}$, $i = 1, 2, ..., n_V$
$v_i^{(t)} \sim p(v_i^{(t)} = 1|H^{(t-1)})$;
(5) When $t < T$, process step back to (2);
(6) When $t = T$, finish the sampling and set $V' = V^T$;

According to Eq.(6) and (7), the probability distribution of sampling for each node in the algorithm can be calculated. After many iterations sampling, v' $v^{(T)}t$ is obtained, and Gibbs sampling method can ensure the stability of the sampling results. According to the previous research, the RBM model can be sampled only once ($T = 1$), when $T > 1$ the performance of the Algorithm is not improved obviously.

After Gibbs sampling, the state of H and $V^{'}$ can be obtained, and the row vector of the i row of W representing the value matrix by $W_{i,*}$, and The parameters update formula for RBM is as follows:

$$W_{i,*} = W_{i,*} + \alpha[p(h_i = 1|V^{(0)})V^{(0)} - p(h_i = 1|V^{'})V^{'}]) \tag{8}$$

$$a = a + \alpha[V^{(0)} - V^{'}]) \tag{9}$$

$$b = b + \alpha[p(H = 1|V^{(0)}) - p(H = 1|V^{'})]) \tag{10}$$

$$p(H=1|V^{(t)}) = [p(h_1=1|V^{(t)}), p(h_2=1|V^{(t)}), \ldots, p(h_H=1|V^{(t)})]^T \tag{11}$$

The learning rate is α, and integrating the Gibbs sampling and parameter updating formulas, and the entire RBM training algorithm is as follows:

Algorithm 2 Training algorithm flow

(1) Initialize the RBM parameters, in general, initialize the Value Matrix $W \sim N(0, 0.01)$, visible layer and hidden layer set a=b=0;
(2) Input individual or batch training samples into RBM for Gibbs sampling;
(3) Update the RBM parameters according to Eq.(7) to (10);
(4) Calculation of RBM model energy in accordance with Eq.(2),If the energy function is not convergent or does not meet the specified training steps, return to Eq.(2),Otherwise,out of the training;

3.2 The Back Propagation Method Is Used to Fine-tune The Network

The deep confidence network (DBN) is a probability generation model, it can reconstruct the input data based on the probability distribution of the input data. The DBN is stacked by multiple RBM models, usually at the end of which a full-connection layer is added as network output and the network structure is shown in figure 6. Each RBM in DBN can be used as a feature detector to extract features from the upper layer, and multiple RBM layers stack up to form a neural network of sufficient depth, and finally, the depth feature of the original input is exported in the last RBM.

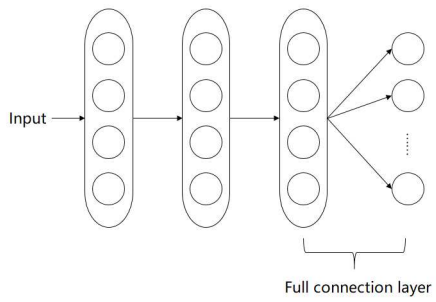


Fig. 6. DBN network structure

The difference between DBN and traditional neural network is that the training process of DBN is divided into two stages: pre-training and fine-tuning.

Pre-training stage: train the network parameters layer by layer by using the RBM model property. When the energy of the layer is stable, the output of this layer is used as the input of the next layer to train the next layer of RBM's parameters. The pre-training is terminated until the last full connection layer is reached. So far the network parameters obtained in the pre-training have been initialized to a better value, and the RBM models have converged to the distribution of input data as much as possible.

Fine-tuning stage: the number of DBN network outputs is equal to the number of classifications when used for classification purposes. Fine-tune the steps using the Back Propagation, combined with mini-batch, to fine tune the overall network parameters. The last layer uses softmax as output, and the network loss function uses the following cross entropy:

$$Loss(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(h_\theta(x^{(i)})) \tag{12}$$

In this equation,θ is DBN network parameters, $x^{(i)}$ is the input data,$y^{(i)}$ is the classification label for the corresponding input, $h_\theta(x^{(i)}$ is the output of the network. The DBN training process is as follows:

Algorithm 3 Sampling algorithm flow

Require: $(X, Y) = [(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_m, y_m)]$,$x_i$ is the network input,y_i is the actual classification of the corresponding input.
(1) In the pre-training stage, let DBN model have K-layer RBM, and $[x_1, x_2, x_3, \ldots, x_m]$ as the input of the first layer of RBM, and in accordance with the training methods in section 2.1, training the first layer of RBM until the energy value of the energy function converges, and makes $[x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_m^{(1)}]$ the output of the first layer of RBM.
(2) Let $k = 2$,using $[x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_m^{(k-1)}]$ as the input training RBM of the k-layer RBM until the energy function converges, and get $[x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_m^{(k)}]$. So on, training the RBM layer by layer in hierarchy order.
(3) In the fine-tuning stage, X is input to the DBN to get the prediction $\widetilde{Y}$ and the crossover entropy is used to calculate the error of classification, and using Back Propagation to adjust the DBN network parameters.

3.3 Using Deep Confidence Network in DTI Prediction

The interactions between drugs and proteins depend to a large extent on their chemical structure.Carrier proteins have specificity. They can only transport one kind of substance or similar nature when they transmit materials inside and outside the cell, and different affinity for different substances is partly determined by the structure of drugs and protein [CLH+13].

In the DTI prediction, we need to determine whether drugs and proteins interact with each other. This is actually a dichotomy problem in which drugs and proteins are combined to form a drug-protein pair. It is divided into two categories: Interaction and non-interaction. DBN is usually used to solve classification problems. It uses the object's signature vector as input, and the final output classification label. In DTI prediction, feature vectors of drug and protein can be calculated separately, they then combine the two groups as data vectors for the DBN. The following is a description of the feature vector construction of drugs and proteins:

3.3.1 Drug's Feature

Drugs are made up of a variety of compounds, and different compounds have different structural, physical properties and chemical property, and each compound corresponds to International Chemical Identifier(InChI) and simplified molecular input line entry specification (SMILES). SMILES can be used to describe molecular expressions and the three-dimensional chemical structure of molecules, as shown in figure 7, but SMILES is not fixed and is not a binary valued, so it is not a suitable input for DBN. Drug fingerprints are commonly used to describe the structure of a compound.

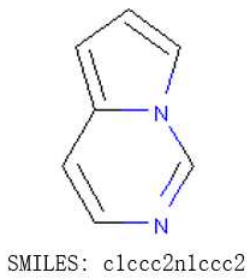


Fig. 7. The SMILES of the compounds and its two-dimensional structure

Common fingerprints are: topological fingerprints, MACCS and Morgan fingerprints, and this paper uses the extended connectivity fingerprint (ECFPs) in the topological fingerprint as the feature vector of the compound. ECFPs is a circular topological fingerprint which is widely used in molecular description and similarity calculation and is widely used in the field of drug research [RH10]. Each of the substructures in ECFPs has a single identifier, and it can use a string of fixed binary sequences to record each substructure,each bit indicating whether there is a substructure.Its expression is suitable for the use of DBN. Figure 8 shows the substructure and its identification of each atom of the same molecule in different neighborhood radius.

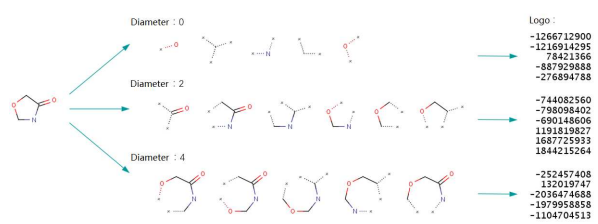


Fig. 8. Substructure identification in a circular neighborhood of different diameters

3.3.2 Protein's Feature

Similar to the SMILES of medicinal, the characteristics of proteins need to be described by using a descriptor, which is described in this paper using an amino acid sequence descriptor. The characteristic length of the n-element amino acid descriptor is 20^n . The larger n is, the more primitive the protein imformation is, but the characteristic length will increases exponentially. So the three-peptide structural characteristics(TC) is used to describe the characteristic of the protein. This method describes the frequency of three consecutive amino acid sequences in a protein sequence. The characteristic length of the protein generated by the descriptor is 8000.

3.3.3 DBN Network Structure

Through the above description of drug and protein feature, combined with the feature of drug and protein, can form a feature vector of 10048 in length. As shown in figure 9, the first 2048 are drug feature, and the latter 8,000 are protein feature. The DBN network structure used in this paper is shown in figure 10. And the paper use three-layer RBM model. The number of output units is 1. The number of units from input to output is 10048-500-500-500-1, and all activation function is sigmoid function.

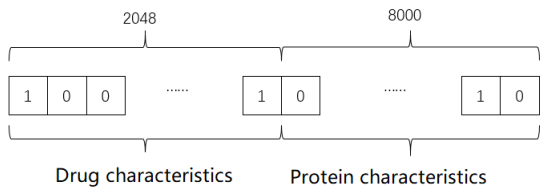


Fig. 9. The drug-protein pair's feature vector

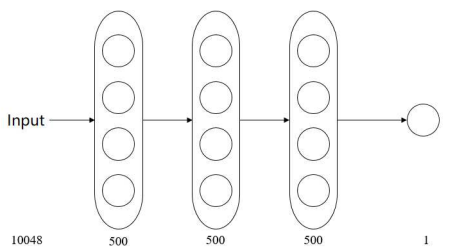


Fig. 10. DBN network structure for DTI prediction

Set the network output to $output$ and the predicted threshold t, and the DTI implementation flow chart, as shown in figure 11, requires a training set and validation set to train and validate the DBN. Once the training is complete, it can be used in a predictive framework. Prior to the prediction phase, drug and protein information is collected from publicly available drug databases such as Drugbank, Chembl and Kegg, that means need to build the SMILES of medicines and protein amino acid sequences. In the prediction phase, you also need to enter the corresponding drug and protein feature sequence,and enter them to the DBN model. Finally, the network output is a real number on domain$[0, 1]$, and compares the real value with the threshold t.When the real number is greater than t, the prediction result is that a given drug interacts with the protein and, conversely, no interaction occurs.

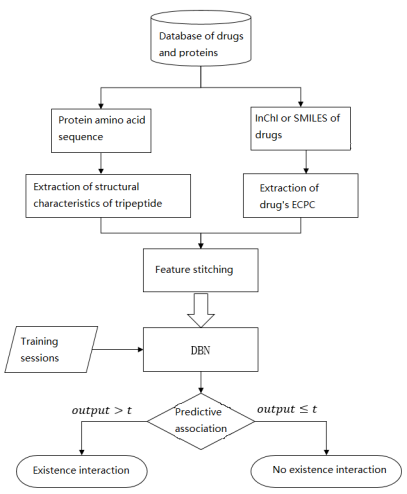


Fig. 11. Process framework for DTI prediction using DBN

Table 1. The basic datasets of the experimental comparison in this paper

Data sets	Number of drugs	Number of protein	Incidence number
Enzyme	445	664	2926
GPCRs	223	95	635
Ion channel	210	204	1476
Nuclear receptor	54	26	90

Table 2. The datasets based on Drugbank

Data sets	Number of drugs	Number of protein	Incidence number
approve	1802	1705	6922
approve+experiment	742	478	5211

4 Analysis of Experiments And Results

Based on four standard data sets, this paper simulates DTI prediction algorithm based on implicit semantic model, network and deep learning, and AUPR and AUROC evaluation indicator are used to compare these advantages and disadvantages and make specific analysis. Then, the self-built large data set based on $Drugbank$ database is used to analyze the predictive performance of DBN based DTI prediction for large-scale data.

4.1 Experimental Data Sets

In this paper, four data sets of $atkinson$, $GPCRs$, $Ionchannel$ and $Nuclearreceptor$ proposed by Yamanishi were used for experiments, as shown in table 1:

In the experiment, five fold cross validation was used to divide four data sets into five equal parts, and the experimental results were the average value of five experiments. In addition, two larger datasets have been constructed based on the $approved$ and $experiment$ datasets in the $Drugbank$ database. The $approved$ datasets have been approved by the FDA. Experiment datasets contain drugs and proteins that have interacted with each other in past experiments but have not been approved by the FDA. As shown in table 2, $approved$ represents an FDA-approved dataset. In this dataset, cold associations are not processed. $approvedexperimentis$ a combination of FDA certification and experimental certification, and the dataset excludes cold associations data, which drugs and proteins with a correlation of less than 2. The two datasets randomly selected 20% of the association as the test set and the remaining 80% as a training set.

The negative sample selection in the dataset is sampled at random and the drug-protein pair with the same number of positive samples is selected as a negative sample, the total association size of $Enzyme$, $GPCRs$, $Ionchannel$, $Nuclearreceptor$, $approved$ and $approved + experiment$ is 5852,1270,2952,180,13844 and 10422.

Feature extraction of drugs and proteins using open source software rdkit and propy.Rdkit is a software for processing chemical informatics, which can quickly obtain chemical structure information through the compounds' InChI or SMILES, and has a large number of built-in functions to calculate certain properties and fingerprint characteristics of the compound. Propy is used to calculate the feature of amino acid sequence of proteins, and it has a lot of built-in protein descriptors [CXL13]. Both of these tools have corresponding libraries on python, allowing quickly get the feature of drug and protein, and then stitching together the two features to form the input vector of the DBN.

Table 3. The division of the results under the classification problem

Actual classification \ Forecast classification	Positive sample	Negative sample
Positive sample	TP	FN
Negative sample	TP	FN

4.2 Using Deep Confidence Network in DTI Prediction

In this paper, we use the following baseline methods, which belong to the implicit semantic model, and the combination of network deep learning and SVM.

(1)Convex hull non-negative matrix factorization(CHNMF):CHNMF is a method of matrix decomposition. In this paper, after comparing the performance effects of the DTI prediction performance under four data sets, 5 kinds of matrix factorization, such as non-negative matrix factorization and binary matrix factorization, are compared, and CHNMF method with the best average performance under these four data sets is selected. This method uses only the drug-protein association matrix, and the other information is not involved in matrix factorization.

(2)Weighted distribution method(WM) :WM is a network-based DTI prediction method proposed by Yamanishi. It uses drug similarity, protein similarity and known drug protein interactions to make DTI predictions, in which drug similarity uses the SIMCOMP method to calculate the structural similarity between drugs. The protein similarity was calculated by using the Smith-Waterman fraction to calculate the similarity of Amino acids between proteins, and it takes similarity as weight, weighted to calculate the predicted interactions between the drug and the protein.

(3)Multiscale feature depth representation interaction prediction(MFDR) :MFDR is a method of combining deep learning and SVM. It first uses feature vector training SAE, then compresses the original feature vector coding through SAE, then input to SVM training, and finally get a binary classifier. The SVM kernel function is set to a radial basis function.

4.3 Using Deep Confidence Network in DTI Prediction

This paper mainly uses AUPR and AUROC as the evaluation index of the Algorithm. The classification results can be divided into four parts in the classification problem, as shown in table 3. According to these four parts, we can get the four evaluation indexes of precision rate($Precision$), recall rate($Recall$), true positive rate(TPR) and false positive rate(FPR), namely formula (13)-(15).

$Precision$ and $Recall$ under the same threshold value are set as a group, and using $Recall$ as the abscissa and $Precision$ as the ordinate. By connecting each point in order, a curve can be obtained, which is called the PR curve. In the same way, the ROC curve is obtained by using the FPR as the abscissa and TPR as the ordinate. The following areas of the two curves are AUPR and AUROC respectively. The ROC curve is a commonly used evaluation index in medical research.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{14}$$

$$FPR = \frac{FP}{TN + FP} \tag{15}$$

The precision rate and recall rate are two contradictory indexes, and the PR curve can be used to measure the equilibrium between the two. The ROC curve shows the relationship between true positive rate and false positive rate, it is helpful to verify the recognition ability of the classifier

Table 4. Comparison of AUPR values of each algorithm

	Enzyme	GPCRs	Ion channel	Nuclear receptor
CHNMF	0.7769	0.7998	0.7360	0.7528
WM	0.8802	0.8681	0.7628	0.6868
MFDR	0.6818	0.8307	0.7569	0.7400
DBN	0.8902	0.8374	0.8082	0.7558

Table 5. Comparison of AUROC values of each algorithm

	Enzyme	GPCRs	Ion channel	Nuclear receptor
CHNMF	0.7368	0.7296	0.6839	0.7253
WM	0.8207	0.8511	0.7372	0.6111
MFDR	0.7203	0.8263	0.7676	0.7809
DBN	0.9006	0.8420	0.8270	0.6821

at different thresholds. The larger the area under PR curve and ROC curve is, the better the performance of classifier is.

4.4 Experimental Results And Analysis

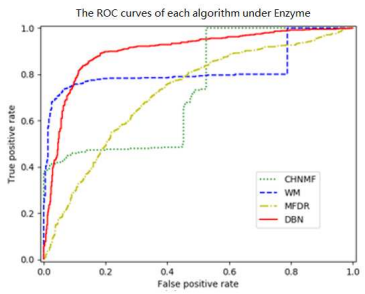
4.4.1 Experiment 1: Comparison of The Prediction Experiments of Each New Method in Table 1 Data Set

The AUPR values and AUROC values of the algorithms under different data sets are given in tables 4 and 5 respectively. As can be seen from the table, the DBN method used in this paper has better AUPR values in $Enzyme$, $Ionchannel$ and $Nuclearreceptor$ than the other three methods, and AUROC values are superior to other methods in the $Enzyme$ and $Ionchannel$.

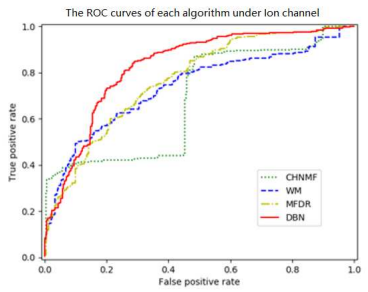
From the above table analysis, it is found that DBN has poor prediction performance under training of small data. As the size of data set increases, DTI prediction performance based on DBN also improves. This indicates that DTI prediction method based on DBN is more suitable for larger drug and protein samples, and not for the insufficient data.

Compared with the MFDR method using SAE in deep learning and SVM, it can be seen from the table that DBN is superior to MFDR in the other three data sets except the performance of $Nuclearreceptor$ is lower than that of MFDR. In addition, MFDR performs worse as the data gets larger. This may be due to the fact that the SAE in MFDR did not fully learn this sample space during the learning phase. In $Nuclearreceptor$, SAE was able to fully learn the overall sample space in $Nuclearreceptor$ due to the small number of nuclear receptor proteins. But in the case of $Enzyme$, due to the fact that there are too many types of enzyme to be fully learned, the coding output of SAE fails to correctly represent the original features, which affects the learning and training of SVM.

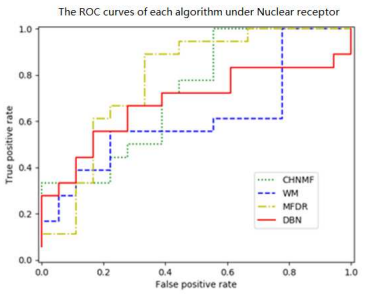
DBN, by means of the RBM structure, trains the network parameters of each layer in order of hierarchy, makes the network fit the distribution of training set as best as possible, and finally trains the whole classification by the method of fine-tuning. Finally, constructing a classifier which can well express the input features, and the classifier is deeper level than the SVM. This allows DBN to perform better in big data sets, while under small data sets, there is a poor performance due to too little data fitting. Figure 14 compares the ROC curves of each algorithm in different data sets.



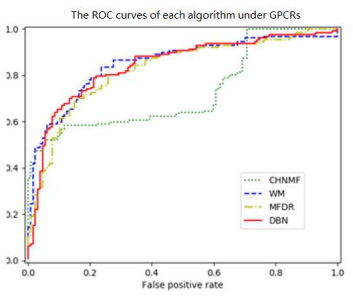
(a) Enzyme



(b) Ion channel



(c) Nuclear recptor



(d) GPCRs

Fig. 12. The ROC curve of each algorithm in each data set

In figure 12, you can see: (1)Because of the small amount of $Nuclearreceptors$ datasets, the ROC curve of the four algorithms is in a ladder shape. In the other three data sets, the ROC curve becomes smooth with the increase of data. The effect of the CHNMF method is that the worst of the four methods, no matter under which the data set shows the shape of a ladder, this shows that the CHNMF can not well divide the true positive samples and the true negative samples, and the small changes in the classification threshold will seriously affect the classification prediction effect.

(2)In figure 12(a), the non-deep-learning CHNMF and WM methods, the ROC curve under this data set is not smooth. Under the condition that the true positive rate is basically the same, the false positive rate rises sharply, which shows that the DTI prediction method that is not based on deep learning, may lead to performance degradation due to the increase of sample space and the failure of the algorithm to accurately reflect the relationship between samples. The ROC curve of the MFDR method based on deep learning is smooth, but it is the worst evaluated algorithm in $Enzyme$, either in the AUPR or in the AUROC evaluation index, which may be caused by two factors:1)There are too many samples of drugs and proteins in $Enzyme$ data, and SAE has not been able to correctly encode the characteristics of drugs and proteins;2)SVM is a shallow-layer learning model, which can not be classified correctly in the later classification learning.

(3)Besides $Nuclearreceptors$, the ROC curve of DBN is better than the other three algorithm, which shows that the DBN algorithm can distinguish the positive and negative samples well under the other three data sets. It can be inferred from the above observations that the performance effect of DBN is affected by the scale of experimental data. Generally, the prediction performance of DBN will also be improved when the data scale increases.

4.4.2 Experiment 2: DBN Method in The Prediction Experiment of The Data Sets in Table 2

The network structure of this experiment is the same as that in the previous paper (10048-500-500-500-1). The truncated normal distribution which average value of the initialization network parameters is 0 and the standard deviation is 0.01 is used as the output random number. In the pre-training stage, the learning rate is 0.001, and the first three layers of RBM are pre-trained. Gibbs sampling iteration times are 1, batch size is 32, and the pre-training times of each layer are 800,1000,1000 respectively. In the fine-tuning stage, the initial learning rate is 0.0001, and the learning rate is automatically adjusted according to the change of cross entropy. $batch$ size is 16, and a total of 10,000 iterations are performed.

Figure 15 shows the value of AUPR and AUROC recorded for DBN in each iteration 200 times in the fine-tuning stage after pre-training under two data sets. Since there is no pre-training in the last layer, both AUPR and AUROC are around 0.5 at the beginning of fine-tuning.

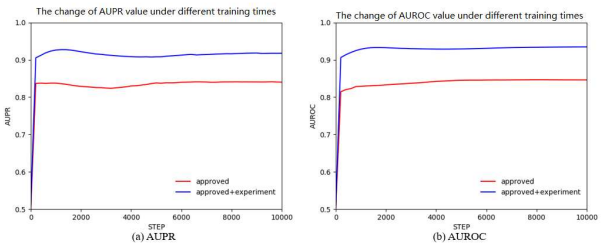


Fig. 13. The AUPR and AUROC curves of DBN under different iterations

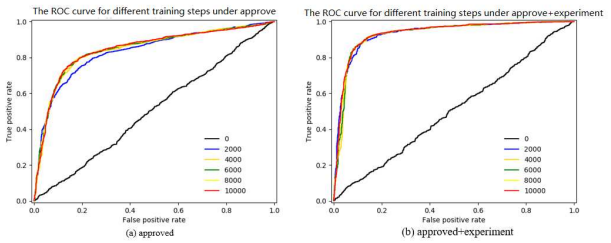


Fig. 14. The ROC curve of different iterations of DBN under the $approved + experiment$ data set

In figure 13 and 14, you can see:

(1)After pre-trained Network, and after about 400 times during the fine-tuning phase,AUPR and AUROC in $approved$ increased to 0.83762 and 0.8203 respectively,and and AUROC in the $approved + experiment$ increased to 0.9120 and 0.91368 respectively. That means after pre-training, DBN can already get a better result after about 400 fine-tuning iterations.By the training of 4000 steps, AUROC was basically stable, while AUPR was still fluctuating, and it started to be stable about the iteration of 5000 steps.After 10,000 fine-tuning iterations, AUPR and AUROC in $approved$ are stable at 0.8403 and 0.8461 respectively, and AUPR and AUROC in $approved+experiment$ are stable at 0.9174 and 0.9347 respectively.

(2)The experimental results of $approved + experiment$ were all better than $approved$, with a difference of nearly 0.1.The ROC curve at different iterations can also be found that the latter is relatively similar between ROC curves after 2000 iterations. However, there is a gap between the ROC curve of the former in 2000 iterations and the comparison of the latter. This suggests that the convergence rate in the $approved + experiment$ dataset is faster because all the correlations in the dataset are greater than 2.

5 Conclusion

This paper presents the prediction of DTI problem with deep confidence network, which is a deep learning model, and has the ability to construct implicit features autonomously, which can excavate the intrinsic association factor effectively. In order to integrate deep confidence network and DTI problem framework, the characteristics of drug fingerprint and amino acid sequence of protein were transformed and extracted, and as input of the model, the classification results were obtained. In order to verify the effectiveness of the proposed approach in DTI prediction, based on the measured data set and compared with the new method, the experimental results show that the deep confidence network method with the increase of data sets, DTI prediction accuracy improved, which helps solve the problem of data set prediction in a larger space. In future research work, there may be two more problems to be solved:

(1)How the implied association feature is effectively extracted and how it determines its weight in the interaction mining. For the research of feature engineering in data mining, the industry already has the deep learning method (DL) and its related software and hardware platform system, and has obtained good application effect in the field of artificial intelligence. How to apply DL to drug-targeted protein prediction will be an important research direction in the future.

(2)Rapid processing of massive data. With the increase of drug and disease subtypes and the rapid growth of the interaction relation, how to deal with the large number of related relationships will also be an important research direction in the future. Map-Reduce (MR) processing framework

is a popular mass data batch processing framework, which can be used to predict the effects of a large number of drugs-target or protein-protein. In future work, the distributed cluster processing method of this algorithm in MR Framework will be studied continuously. In this way, the prediction results can be obtained quickly and the application of this research is advanced. In the research of data feature mining, we will use DL method to mine the hidden feature sets in order to improve the prediction accuracy of drug-protein.

1 Appendix A. Acknowledge

This work was financially supported by Guangdong Natural Science Foundation Project (2018A030313437) and the open fund project of Guangdong Key Laboratory of Big Data Analysis and Processing (2017006, 2017010).

References

[APGF13]Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, 29(16):2004–2008, 2013.

[CLH+13]Murat Can Cobanoglu, Chang Liu, Feizhuo Hu, Zoltaì'n N Oltvai, and Ivet Bahar. Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling*, 53(12):3399–3409, 2013.

[CXL13]Dong-Sheng Cao, Qing-Song Xu, and Yi-Zeng Liang. propy: a tool to generate various modes of chouâŁ™s pseaac. *Bioinformatics*, 29(7):960–962, 2013.

[CY+16]Keith CC Chan, Zhu-Hong You, et al. Large-scale prediction of drug-target interactions from deep representations. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1236–1243. IEEE, 2016.

[CYZ+15]Xing Chen, Chenggang Clarence Yan, Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin, and Yongdong Zhang. Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics*, 17(4):696–712, 2015.

[DTMZ13]Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in bioinformatics*, 15(5):734–747, 2013.

[EIP+13]Dorothea Emig, Alexander Ivliev, Olga Pustovalova, Lee Lancashire, Svetlana Bureeva, Yuri Nikolsky, and Marina Bessarabova. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, 8(4):e60618, 2013.

[EZW+17]Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 14(3):646–656, 2017.

[GBB+11]Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2011.

[Kan02]Minoru Kanehisa. The kegg database. In *âĽˇIn SilicoâĽ™Simulation of Biological Processes: Novartis Foundation Symposium 247*, volume 247, pages 91–103. Wiley Online Library, 2002.

[LHZ17]Hailiang Li, Yongqian Huang, and Zhijun Zhang. An improved faster r-cnn for same object retrieval. *IEEE Access*, 5:13665–13676, 2017.

[LKD+13]Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2013.

[RH10]David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[SMH07]Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.

[WZ16]Fangping Wan and Jianyang Zeng. Deep learning with feature embedding for compound-protein interaction prediction. *bioRxiv*, page 086033, 2016.

[YAG+08]Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[YKKG10]Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010.